Experiential AI

Experiential AI

Drew Hemment (University of Edinburgh; drew.hemment@ed.ac.uk)
Ruth Aylett (Herriot Watt University; r.s.aylett@hw.ac.uk)
Vaishak Belle (University of Edinburgh; Vaishak@ed.ac.uk)
Dave Murray-Rust (University of Edinburgh; D.Murray-Rust@ed.ac.uk)
Ewa Luger (University of Edinburgh; Ewa.Luger@ed.ac.uk)
Jane Hillston (University of Edinburgh; Jane.Hillston@ed.ac.uk)
Michael Rovatsos (University of Edinburgh; Michael.Rovatsos@ed.ac.uk)
Frank Broz (Herriot Watt University; f.broz@hw.ac.uk)
DOI: 10.1145/3320254.3320264

Abstract

Experiential AI is proposed as a new research agenda in which artists and scientists come together to dispel the mystery of algorithms and make their mechanisms vividly apparent. It addresses the challenge of finding novel ways of opening up the field of artificial intelligence to greater transparency and collaboration between human and machine. The hypothesis is that art can mediate between computer code and human comprehension to overcome the limitations of explanations in and for AI systems. Artists can make the boundaries of systems visible and offer novel ways to make the reasoning of AI transparent and decipherable. Beyond this, artistic practice can explore new configurations of humans and algorithms, mapping the terrain of inter-agencies between people and machines. This helps to viscerally understand the complex causal chains in environments with AI components, including questions about what data to collect or who to collect it about, how the algorithms are chosen, commissioned and configured or how humans are conditioned by their participation in algorithmic processes.

Introduction

AI has once again become a major topic of conversation for policymakers in industrial nations and a large section of the public.

In 2017, the UK published Ready, Willing and Able, a landscape report (House Of Lords Select Committee, 2018). It clearly states that "everyone must have access to the opportunities provided by AI" and argues the need for public understanding of, and engagement with AI to develop alongside innovations in the field. The report warns of the very real risk of "societal and regional inequalities emerging as a consequence of the adoption of AI and advances in automation" (*Ibid.*). It also assesses issues of possible harm from malfunctioning AI, and resulting legal liabilities. However, it stops short of considering more pervasive downsides of applying AI decision-making across society. Alongside the sometimes exaggerated claims of AI's current or immediate-future capabilities, a broader set of fears about negative social consequences arise from the fast-paced deployment of AI technologies and a misplaced sense of trust in automated recommendations. While some of these fears may themselves be exaggerated, negative outcomes of ill-designed data-driven machine learning technologies are apparent, for example where new knowledge is formulated on undesirably biased training sets. The notorious case of Google Photos grouping some humans with primates on the basis of skin tone offered a glimpse of the damage that can be done. Such outcomes may not be limited to recommendations on a mobile phone: social robots share everyday spaces with humans, and might also be trained on impoverished datasets. Imagine, for example, a driverless car not recognizing specific humans as objects it must not crash into. So much for Asimovs laws!

Accountability and explainability in AI

The AI community has, of course, not been silent on these issues, and a broad range of solutions have been proposed. We broadly

classify these efforts into two related categories: accountability and explainability.

The first category seeks to identify the technical themes that would make AI trustworthy and accountable. Indeed, we can see AI technologies are already extending the domains of automated decision making into areas where we currently rely on sensitive human judgements. This raises a fundamental issue of democratic accountability, since challenging an automated decision often results in the response "it's what the computer says". So operators of AI need to know the limits and bounds of the system, the way that bias may present in the training data, or we will see more prejudice amplified and translated to inequality. From the viewpoint of AI research, there is a growing scientific literature on fairness (Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018) to protect those otherwise disenfranchised through algorithmic decisions, as well as engineering efforts to expose the limitations of systems. Accountability can be a deeper property of the system too: for example, an emerging area of AI research looks at how ethical AI systems might be designed (Conitzer, Sinnott-Armstrong, Borg, Deng, & Kramer, 2017; Halpern & Kleiman-Weiner, 2018; Hammond & Belle, 2018).

The second category investigates how the decisions and actions of machines can be made explicable to human users (Gunning, 2017). We are seeing a step change in the number of people both currently and potentially impacted by automated decisions. Whilst the use of algorithms can now be said to be common (Domingos, 2015), concerns arise where complex systems are applied in the generation of sensitive social judgments, such as in social welfare, healthcare, criminal justice, and education. This has led to a call to limit the use of "black box" systems in such settings (Campolo, Sanfilippo, Whittaker, & Crawford, 2017). However, if one asks for a rationale for a decision, usually none is given, not least because those working in organisations using automated decision-making do not themselves have any insight into what the algorithms driving it are doing. This is a form of conditioning, creating passivity rather than engagement. At the other extreme, if people do not understand the decisions of AI systems, they may simply not use those systems. Be that as it may, progress in the field has been exciting but a single solution is elusive. Some strands of research focus on using simpler models (possibly at the cost of prediction accuracy), others attempt "local" explanations that identify interpretable patterns in regions of interest (Weld & Bansal, 2018; Ribeiro, Singh, & Guestrin, 2016), while still others attempt human-readable reconstructions of high-dimensional data (Penkov & Ramamoorthy, 2017; Belle, 2017). However, this work addresses explainability as primarily a technical problem, and does not account for human, legal, regulatory or institutional factors. What is more, it does not generate the kind of explanations needed from a human point of view. A person will want to know why there was one decision and not another, the causal chain, not an opaque description of machine logic. There are distinctions to be explored between artificial and augmented intelligences (Carter & Nielsen, 2017), and a science, and an art, to be developed around human-centred machine learning (Fiebrink & Gillies, 2018).

For there to be responsible AI, transparency is vital, and people need comprehensible explanations. Core to this is the notion that unless the operation of a system is visible, and people can access comprehensible explanations, it cannot be held to account. Even when an explanation can be provided, this may not always be sufficient (Edwards & Veale, 2017) and more intuitive solutions are required to, for example, understand the changing relations between data and the world, or integrate domain knowledge in ways that connect managers with those at the frontlines (Veale, Van Kleek, & Binns, 2018). In Seeing without knowing, Ananny and Crawford argue research needs not to look *within* a technical system, but to look *across* systems and to address both human and non-human dimensions (Ananny & Crawford, 2018). They call for "a deeper engagement with the material and ideological realities of contemporary computation" (*Ibid.*).

Artists addressing such AI challenges

There is a mature tradition of work between art and technology innovation going back to the 1960s and 1970s (Harris, 1999; Gere,

Figure 1: Neural Glitch 1540737325 Mario Klingemann 2018

2009). Artists are beginning to experiment in AI as subject and tool, and several high profile programmes are a testament to the fertility of this field (Zentrum fur Kunst und Medien, 2018; Ars Electronica, 2018). Such practice can create experiences around social impacts and consequences of technology, and create insights to feed into the design of the technologies (Hemment, Bletcher, & Coulson, 2017).

One theme evident among artists working with machine learning algorithms today, such as Mario Klingemann[1] and Robbie Barrat[2], is to reveal distortions in the ways algorithms make sense of the world – see Figure 1 for an example. This kind of approach enables the character of machine reasoning and vision to be made explicit, and its artifacts to be made tangible. This, in turn, creates a concrete artefact or representation that can be used as an object for discussion and to spark further enquiry, helping to build literacy in those systems.

In the contemporary experience of AI, the disturbing yet compelling output of DeepDream has shaped our view on what algorithms do, although it is questionable how representative this is of deep network structures, or whether it is a happy accident in machine aesthetics. Either way, it has prompted artistic exploration of the social implications of AI, with projects using deep learning to generate faces (*Plugging 50,000 portraits into facial recognition*, 2018) and Christies auctioning neural network generated portraits (*Is artificial intelligence set to become arts next medium?*, 2018). Going beyond the typical human+computer view, artists are questioning the construction of prejudice and normalcy (*http://mushon.com/tnm*, 2018), and working with AI driven prosthetics, to open possibilities for more intimate entanglements (Donnarumma, 2018).

Art can both make ethical standards concrete, and allow us to imagine other realities. While high-level ethical principles are easy to articulate, they sit at a level of generality that may make their practical requirements less obvious. Equally, they signal the existence of clear solutions, externalise responsibility, and obscure the true complexity of the moral problems resulting from socially situated AI. Ethical issues must be concretely internalised by developers and users alike to avoid failures like Cambridge Analytics or the Facebook Emotional Contagion experiment (Jouhki, Lauk, Penttinen, Sormanen, & Uskali, 2016). Experiential approaches (Kolb, 2014) can act as a powerful mechanism, and embedding relevant experiences in a story-world through narrative, and especially role-play, can generate safe reflection spaces," as for example Boal's Forum Theatre (Boal, 2013).

Accountability is variously addressed. Joy Buolamwini works with verse and code to challenge harmful bias in AI[3], while Trevor Paglen constructs a set of rules for algorithmic systems in such a way as to uncover the character of that rule space[4]. A thriving community of practitioners from across the arts and sciences is working to avoid detection [5] or trick classification systems (Sharif, Bhagavatula, Bauer, & Reiter, 2016). Such artistic experiments bring to life and question what an algorithm does, what a system could be used for, and who is in control.

Experiential AI theme and call for artists

The field of Experiential AI seeks to engage practitioners in computation, science, art and design around an exploration of how humans and artificial intelligences relate, through the physical and digital worlds, through decisions and shaping behaviour, through collaboration and co-creation, through intervening in existing situations and through creating new configurations.

The Experiential AI theme begins with a call for artists in residence, launched in July 2019, as a collaboration between the Experiential AI group at University of Edinburgh, Ars Electronica in Linz, and Edinburgh International Festival [6]. The focus is on creative experiments in which AI scientists and artists are jointly engaged to make artificial intelligence and machine learning tangible, interpretable, and ac-

[1] http://quasimondo.com/
[2] https://robbiebarrat.github.io/
[3] https://www.poetofcode.com/
[4] http://www.paglen.com/
[5] https://cvdazzle.com/
[6] https://efi.ed.ac.uk/art-and-ai-artist-residency-and-research-programme-announced/

cessible to the intervention of a user or audience. The ambition is to help us think differently about how algorithms should be designed, and open possibilities for radically new concepts and paradigms.

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Ars Electronica. (2018). *Media art between natural and artificial intelligence.* https://ars.electronica.art/ai/en/media-art-between-natural-and-artificial-intelligence/.

Belle, V. (2017). Logic meets probability: Towards explainable ai systems for uncertain worlds. In *Ijcai* (pp. 5116–5120).

Boal, A. (2013). *The rainbow of desire: The boal method of theatre and therapy*. Routledge.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). Ai now 2017 report. *AI Now Institute at New York University*.

Carter, S., & Nielsen, M. (2017). Using artificial intelligence to augment human intelligence. *Distill*, *2*(12), e9.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Thirty-first aaai conference on artificial intelligence.*

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Penguin.

Donnarumma, M. (2018). *Is artificial intelligence set to become arts next medium?* https://marcodonnarumma.com/works/ai-ethics-prosthetics/.

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, *16*, 18.

Fiebrink, R., & Gillies, M. (2018, June). Introduction to the special issue on human-centered machine learning. *ACM Trans. Interact. Intell. Syst.*, *8*(2), 7:1–7:7. Retrieved from http://doi.acm.org/10.1145/3205942 doi: 10.1145/3205942

Gere, C. (2009). *Digital culture*. Reaktion Books.

Gunning, D. (2017). *Explainable artificial intelligence (xai).* https://tinyurl.com/yccmn477. (Accessed: 12/3/18)

Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Thirty-second aaai conference on artificial intelligence.*

Hammond, L., & Belle, V. (2018). Deep tractable probabilistic models for moral responsibility. *arXiv preprint arXiv:1810.03736*.

Harris, C. (1999). The xerox palo alto research center artist-in-residence program landscape. In *Art and innovation* (pp. 2–11).

Hemment, D., Bletcher, J., & Coulson, S. (2017). Art, creativity and civic participation in iot and smart city innovation through open prototyping. In *Proceedings of the creativity world forum* (pp. 1–2).

House Of Lords Select Committee. (2018). *Ai in the uk: ready, willing and able?* https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Is artificial intelligence set to become arts next medium? (2018). https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx.

Jouhki, J., Lauk, E., Penttinen, M., Sormanen, N., & Uskali, T. (2016). Facebooks emotional contagion experiment as a challenge to research ethics. *Media and Communication*, *4*.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In *Aea papers and proceedings* (Vol. 108, pp. 22–27).

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.

Penkov, S., & Ramamoorthy, S. (2017). Using program induction to interpret transition system dynamics. *arXiv preprint arXiv:1708.00376*.

Plugging 50,000 portraits into facial recognition. (2018). https://www.reddit.com/r/Damnthatsinteresting/comments/9udese/plugging_50000

_portraits_into_facial/.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security* (pp. 1528–1540).

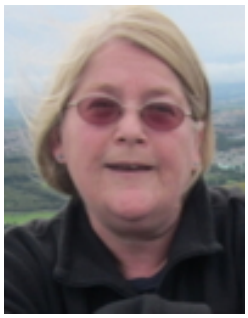
(2018).

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems* (p. 440).

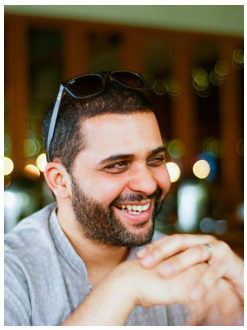
Weld, D. S., & Bansal, G. (2018). Intelligible artificial intelligence. *arXiv preprint arXiv:1803.04263*.

Zentrum fur Kunst und Medien. (2018). *Encoding cultures: Living amongst intelligent machines.* https://zkm.de/en/event/2018/04/encoding-cultures-living-amongst-intelligent-machines.

Drew Hemment is a Chancellors Fellow and Reader at Edinburgh Futures Institute and Edinburgh College of Art, University of Edinburgh. He is PI of of GROW Observatory (EC H2020), Founder of FutureEverything, and on the Editorial Board of Leonardo. His work over 25 years in digital arts and innovation has been recognised by awards including STARTS Prize 2018, Lever Prize 2010 and Prix Ars Electronica 2008.

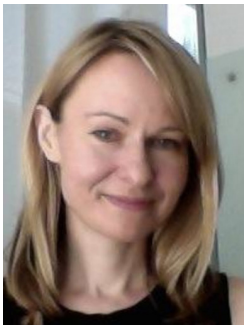


Ruth Aylett is a Professor of Computer Science at Heriot-Watt University in Edinburgh. She researches social agents, Human-Robot Interaction and affective systems, taking a human-centred design approach to the development of socially-useful systems. Her current project explores the use of a robot to assist adults with autism in social signal processing



Vaishak Belle is a Chancellors Fellow/Lecturer at the School of Informatics, University of Edinburgh, an Alan Turing Institute Faculty Fellow, and a member of the RSE (Royal Society of Edinburgh) Young Academy of Scotland. At the University of Edinburgh, he directs the Belle Lab, which specializes in the unification of symbolic systems and machine learning.

Dave Murray-Rust is a Lecturer in Design Informatics at the University of Edinburgh, exploring ways that people, data and things interact. His research centres on how we can ensure that there is space for people within computational systems, preserving privacy, choice, identity and humanity while making use of possibilities of computational coordination and personal data.



Ewa Luger is a Chancellor's Fellow at the University of Edinburgh, a consulting researcher at Microsoft Research UK (AI and Ethics), and a fellow of the Alan Turing Institute. She explores applied ethical issues within the sphere of machine intelligence and data-driven systems. This includes data governance, consent, privacy and how intelligent systems might be made intelligible to the user.

Jane Hillston was appointed Professor of Quantitative Modelling in the School of Informatics at the University of Edinburgh in 2006, having joined the University as a Lecturer in Computer Science in 1995. She is currently the Head of School. Her research is concerned with formal approaches to modelling dynamic behaviour, particularly the use of stochastic process algebras for performance modelling and stochastic verification.

Michael Rovatsos is a Reader (Associate Professor) at the School of Informatics of the University of Edinburgh, and Director of the Bayes Centre for Data Science and AI. He research interests are in intelligent agents and multiagent systems, and most of his recent work has focused on the human-centric of AI algorithms and ethical AI more generally.

Frank Broz is an Assistant Professor of Computer Science at Heriot-Watt University. His research interests are in human-robot interaction, AI, and social robotics. He has consulted for artists working with technology and collaborated on robotic art installations such as Reach, Robot (created for Pittsburghs 250th anniversary celebration). He received his PhD from Carnegie Mellon Universitys Robotics Institute.